

METHOD, DEVICE AND SYSTEM FOR BUILDING A DATASET FOR AUTOMATIC SPEECH RECOGNITION

TECHNICAL FIELD

[0001] Various aspects of this disclosure relate to methods, devices, and systems for building datasets for automatic speech recognition.

BACKGROUND

[0002] Automatic speech recognition (ASR) technology is used to convert spoken language, typically captured as audio signals or data, into written text. ASR technology may be used for transcribing audio files/datasets and performing post-call reviews, amongst others.

[0003] Notwithstanding the increased adoption of ASR technology, current ASR technology may be inadequate to cope with relatively less commonly known languages, for example, Southeast Asian languages, and dialects. In addition, although publicly available data on some of the less commonly known languages exist, these publicly available data may be insufficient for building a sizable dataset for robust ASR. The result of ASR-based recognition may further be affected by the way the audio data is collected, i.e. via scripted reading or spontaneous reading, which may make the results of ASR less consistent or different, and any infrequent vocabulary used may affect the accuracy of the ASR.

[0004] Accordingly, there exists a need for an improved automatic speech recognition system, method, and/or device.

SUMMARY

[0005] The technical solution seeks to provide a method, device, and/or system for building one or more datasets for automatic speech recognition. In some aspects the dataset forms at least part of a training dataset for training an artificial intelligence-based automatic speech recognition (ASR) module to recognize relatively uncommon languages. In some aspects, the technical solution may comprise a device for generating input dataset for speaker verification, and obtaining audio datasets from one or more selected users. The obtained audio datasets may be assigned pseudo labels to facilitate the semi-supervised training of the ASR module.

[0006] In an aspect of the present disclosure, there is provided a computer-implemented method for building a dataset for automatic speech recognition, the method comprising: generating an input dataset; obtaining an audio dataset associated with, corresponding to or based on the input dataset; verifying the audio dataset using an audio verification module; assigning at least one pseudo-label to the verified audio dataset; storing the at least one pseudo-labelled audio dataset in a training database; repeating the above steps until a pre-determined number of pseudo-labelled audio datasets have been stored in the training database; and training an automatic speech recognition module based on the pre-determined number of pseudo-labelled audio datasets.

[0007] In some embodiments, the input dataset may comprise a text dataset, and generating the text dataset comprises introducing at least one of a relatively uncommon vocabulary and a relatively uncommon language in the text dataset.

[0008] In some embodiments, the method further comprises comparing the text dataset for speaker verification with an output transcription file of the automatic speech recognition module, and identifying at least one unsatisfactory result from the comparison. In some embodiments, the step of identifying the at least one unsatisfactory trained result may include determining a precision rate and/or a recall rate associated with one or more words in the text dataset.

[0009] In some embodiments, the method further comprises a step of selecting a speaker from a user group historically associated with accurate text-dependent speaker verification, and assigning the at least one pseudo-label to the audio dataset generated by the speaker without verification.

[0010] In some embodiments, the method further comprises generating an output transcription file by the automatic speech recognition module.

[0011] In some embodiments, the input dataset comprises a first text dataset and a second text dataset, resulting in a respective first output transcription file and a second output transcription file, wherein the first text dataset and the second output transcription file is configured to be input to a first inference module, and the second text dataset and the first output transcription file is configured to be input to a second inference module.

[0012] In some embodiments, the output of first inference module and the output of the second inference module are compared.

[0013] In another aspect of the present disclosure, there is a device for building a dataset for automatic speech recognition, the device comprises a processor configured to repeatedly: generate an input dataset for speaker verification; obtain an audio dataset based on the generated input dataset; verify the audio dataset using an audio verification module; assign at least one pseudo-label to the verified audio dataset; and store the at least one pseudo-labelled audio dataset in a training database until a pre-determined number of pseudo-labelled audio datasets have been stored in the training database.

[0014] In some embodiments, the input dataset comprises a text dataset, the processor is configured to introduce at least one of a relatively uncommon vocabulary and a relatively uncommon language in the generation of the text dataset.

[0015] In some embodiments, the processor is configured to compare the text dataset for speaker verification with an output transcription file of the automatic speech recognition module, and identify at least one unsatisfactory result from the comparison.

[0016] In some embodiments, the processor is configured to identify the at least one unsatisfactory trained result based on a precision rate and/or a recall rate associated with one or more words in the output transcription file.

[0017] In some embodiments, the processor is configured to select a speaker from a user group historically associated with accurate text-dependent speaker verification, and the processor is further configured to assign the at least one pseudo-label to the audio dataset generated by the speaker, without verification.

[0018] In some embodiments, the processor is configured to generate the text dataset based on feedback of the at least one unsatisfactory result.

[0019] In some embodiments, the input dataset comprises a first text dataset and a second text dataset, resulting in a first output transcription file and a second output transcription file respectively, wherein the first text dataset and the second output transcription file is configured to be input to a first inference module, and the second text dataset and the first output transcription file is configured to be input to a second inference module.

[0020] In some embodiments, the output of first inference module and the output of the second inference module are compared.

[0021] In another aspect of the disclosure, there is provided a non-transitory computer-readable storage medium comprising instructions, which, when executed by one or more

processors, cause the execution of the method for building a dataset for automatic speech recognition according to any one of the method embodiments described above.

[0022] In another aspect of the disclosure, there is provided a data processing device configured to perform the method according to any one of the method embodiments described above.

[0023] In another aspect of the disclosure, there is provided a computer executable code comprising instructions for perform the method according to any one of the method embodiments described above.

[0024] In another aspect of the disclosure, there is provided an automatic speech recognition module or system trained by any one of the method embodiments described above.

BRIEF DESCRIPTION OF THE DRAWINGS

[0025] The disclosure will be better understood with reference to the detailed description when considered in conjunction with the non-limiting examples and the accompanying drawings, in which:

- FIG. 1 is a flowchart of a method for building data for automatic speech recognition in accordance with various embodiments;
- FIG. 2 is a block diagram of a system for building datasets for training an automatic speech recognition module in accordance with various embodiments;
- FIG. 3 is a block diagram of a system incorporating cross-validation to improve the quality of pseudo labels assigned to datasets for training an automatic speech recognition module;
- FIG. 4 illustrates an application of the trained ASR module in the form of a post-call analysis system;
- FIG. 5 illustrates an application of the trained ASR module in the form of a transcription service for a vehicle in-trip audio recording; and
- FIG. 6 shows a server computer according to an embodiment.

DETAILED DESCRIPTION

[0026] The following detailed description refers to the accompanying drawings that show, by way of illustration, specific details and embodiments in which the disclosure may be

practiced. These embodiments are described in sufficient detail to enable those skilled in the art to practice the disclosure. Other embodiments may be utilized and structural, and logical changes may be made without departing from the scope of the disclosure. The various embodiments are not necessarily mutually exclusive, as some embodiments can be combined with one or more other embodiments to form new embodiments.

[0027] Embodiments described in the context of one of the enclosure systems, devices, or methods are analogously valid for the other systems, devices, or methods. Similarly, embodiments described in the context of a system are analogously valid for a device or a method, and vice-versa.

[0028] Features that are described in the context of an embodiment may correspondingly be applicable to the same or similar features in the other embodiments. Features that are described in the context of an embodiment may correspondingly be applicable to the other embodiments, even if not explicitly described in these other embodiments. Furthermore, additions and/or combinations and/or alternatives as described for a feature in the context of an embodiment may correspondingly be applicable to the same or similar feature in the other embodiments.

[0029] In the context of various embodiments, the articles “a”, “an” and “the” as used with regard to a feature or element include a reference to one or more of the features or elements.

[0030] As used herein, the term “and/or” includes any and all combinations of one or more of the associated listed items.

[0031] As used herein, the term “dataset” may be understood to include information in any suitable analogue or digital form, for example, provided as a file, a portion of a file, a set of files, a signal or stream, a portion of a signal or stream, a set of signals or streams, waveforms, and the like. The term dataset, however, is not limited to the aforementioned examples and may take various forms and represent any information as understood in the art.

[0032] As used herein, the term “speaker verification” broadly includes a method or system to accept or reject an identity claim by comparing at least two audio samples: one that is used as a reference of the identity and the other that is collected during the test/authentication from the person who makes the claim.

[0033] As used herein, the term “pseudo label or pseudo labelling” broadly includes a process or system of using a labelled data model to predict labels for unlabelled data. At the outset, an initial set of pseudo-labelled data may be generated, this is assumed to be accurate

based on historical records. The initial set of pseudo-labelled data may be used to generate subsequent pseudo-labels for the unlabelled dataset.

[0034] As used herein, the term “module” refers to, forms part of, or includes an Application Specific Integrated Circuit (ASIC); an electronic circuit; a combinational logic circuit; a field programmable gate array (FPGA); a processor (shared, dedicated, or group) that executes code; other suitable hardware components that provide the described functionality; or a combination of some or all of the above, such as in a system-on-chip. The term module may include memory (shared, dedicated, or group) that stores code executed by the processor. A single module or a combination of modules may be regarded as a device.

[0035] As used herein, the term “associate”, “associated”, and “associating” indicate a defined relationship (or cross-reference) between two items. For instance, an audio file can be associated with a text file, indicating that the audio file may be derived, computed, and/or produced using the text file as a source or reference.

[0036] As used herein, “memory” may be understood as a non-transitory computer-readable medium in which data or information can be stored for retrieval. References to “memory” included herein may thus be understood as referring to volatile or non-volatile memory, including random access memory (“RAM”), read-only memory (“ROM”), flash memory, solid-state storage, magnetic tape, hard disk drive, optical drive, etc., or any combination thereof. Furthermore, it is appreciated that registers, shift registers, processor registers, data buffers, etc., are also embraced herein by the term memory. It is appreciated that a single component referred to as “memory” or “a memory” may be composed of more than one different type of memory, and thus may refer to a collective component including one or more types of memory. It is readily understood that any single memory component may be separated into multiple collectively equivalent memory components, and vice versa. Furthermore, while memory may be depicted as separate from one or more other components (such as in the drawings), it is understood that memory may be integrated within another component, such as on a common integrated chip.

[0037] According to an aspect of the disclosure and referring to FIG. 1, there is a method 100 for building a dataset for automatic speech recognition. The method 100 may be implemented by a computer and comprises the steps of generating a input dataset (step 102); obtaining an audio dataset associated with the input dataset (step 104); verifying the audio dataset using an audio verification module (step 106); assigning at least one pseudo-label to

the verified audio dataset (step 108); storing the at least one pseudo-labelled audio dataset in a training database (step 110); repeating steps 102 to 110 until a pre-determined number of pseudo-labelled audio datasets have been stored in the training database (step 112); and training an automatic speech recognition module based on the pre-determined number of pseudo-labelled audio datasets (step 114).

[0038] The input dataset of step 102 may be a text file or text dataset that comprises a plurality of words in a language for verification by a speaker. The plurality of words may be arranged in a sentence “The quick brown fox jumps over the lazy dog”, or as discrete words representing numbers in successive order (e.g. “one”, “two”, “three”...), or as discrete words representing numbers in randomized order (e.g. “nine”, “six”, “one”, “seven”...). In some embodiments, generating the text dataset for speaker verification further comprises introducing at least one of a relatively uncommon vocabulary and a relatively uncommon language in the text dataset. For example, the text file includes common words such as “the”, “in”, and “and”. As another example, the text file includes uncommon words such as “unverified”, “abstain”, “suicide”, etc. Such uncommon words may also include slangs or informal words in a dialect, or may include a mixture of different languages, such as Singlish. In some embodiments, the text file may include Singlish words such as “alamak”, “boleh”, “cannot-lah”, etc. In some embodiments, the input dataset may include non-text dataset and text dataset. In some embodiments, the input dataset may include multimedia content where a text dataset can be extracted therefrom.

[0039] In some embodiments, the introduction of at least one of a relatively uncommon vocabulary and a relatively uncommon language in the text dataset may be performed after a pre-determined number of iterations associated with training the ASR module.

[0040] In some embodiments, the introduction of at least one of a relatively uncommon vocabulary and a relatively uncommon language in the text dataset may be performed after receiving feedback that the ASR module has not been properly trained to transcribe the relatively uncommon vocabulary and/or the relatively uncommon language. The feedback may be generated based on identification of words having a lower precision or recall rates after training the ASR module for some time.

[0041] In some embodiments, a speaker to produce an audio file or dataset associated with the text file for verification is selected from a group of registered users of an audio verification system. In some embodiments, the audio verification system may be a text-dependent speaker

verification system, the group of registered users has previously performed verification using the speaker verification module. The registered users may previously have their audio files enrolled with the speaker verification system, and one or more sub-groups of registered users who are already accustomed to this process and consistently supplied accurate text in a past period during verification may be identified. The audio files previously deemed accurate over a period will form a historical database that is assumed to be accurate/correct for the training of the automatic speech recognition system of the present disclosure. The one or more sub-groups of registered users will be referred to or known as high-quality user sub-group.

[0042] In step 104, a speaker is selected from the high-quality user sub-group to obtain an audio file corresponding to the text dataset.

[0043] In step 106, the verification may be performed by the audio verification module in the form of a text-dependent speaker verification module the speaker is registered with.

[0044] In step 108, each of the verified audio files provided by the speaker selected from the high-quality user sub-group is assigned at least one pseudo-label. The pseudo-label may be in the form of a tag to indicate that the audio file is accurate and will be used for subsequent training of the automatic speech recognition system of the present disclosure. In some embodiments, the pseudo-label may include meta-data associated with the speaker and/or the text file.

[0045] In step 110, the pseudo-labelled audio file may be stored as an entry in a training database. The steps 102 to 108 may then be repeated until a pre-determined number of entries have been entered into the training database.

[0046] In step 112, the pseudo-labelled audio files may be fed into an automatic speech recognition (ASR) model to train the same. The ASR model may include a Machine Learning (ML) algorithm or Artificial Intelligence (AI) algorithm to process human speech into readable text. In some embodiments, the ML/AI algorithm may be trained using unsupervised learning, supervised learning, and/or hybrid training methodology. In some embodiments, the ASR module includes an artificial intelligence module, which may include one or more neural network algorithms.

[0047] In some embodiments, the input dataset to be fed into the ASR may comprise audio files, and the output produced by the ASR may be a text output dataset based on the audio files.

[0048] In some embodiments, the text output dataset may be compared with a testing/test set to identify relatively poor-performing words that may require the collection of more data

for retraining the ASR. The test set may be a separate set from the pseudo-labelled training set and may be pseudo-labelled or properly labelled. In addition, or as an alternative, the text output dataset may be reviewed by one or more users to identify the relatively poor-performing words that may require retraining.

[0049] In another embodiment, the steps 102 to 106 may be replaced by selecting a set of audio files from the high-quality users. The assignment of pseudo-labels is based on the assumption that the high-quality users produces accurate audio files without the need for further verification.

[0050] FIG. 2 shows a system 200 for building a dataset for automatic speech recognition. The system 200 may form part of a device for building datasets for automatic speech recognition. The system 200 comprises a selector module 202 configured to generate input data, such as text data for speaker verification, a collector module 204 configured to obtain an audio file based on the generated text data; an ASR module 206 arranged in data communication with the selector module 202 and the collector module 204. The system 200 may include a review module 208.

[0051] The selector module 202 is configured to generate a text dataset (e.g. words) for speaker verification and subsequent training of the ASR module 206. The selector module 202 may be configured to include uncommon or rare words for each training or retraining of the ASR module 206 so that over time, the ASR module 206 may be trained to recognize more words, including uncommon vocabulary and words in uncommon languages. In some embodiments, the selector module 202 may be configured to introduce words in different languages, such as south-east Asian languages. Non-limiting examples include Thai, Bahasa Malayu, Bahasa Indonesia, Tagalog, etc.

[0052] The collector module 204 may be configured to collect a plurality of pseudo-labelled audio files, and store the plurality of pseudo-labelled audio files in a training database. Multiple pseudo-labelled audio files may be collected until a pre-determined number (target number) of audio files have been collected. The collector module 204 may include a verification module configured to verify the audio file based on the text and assign at least one pseudo-label to the verified audio file.

[0053] In some embodiments, the collected audio files may form a training set to train the ASR module 206. In some embodiments, the collected audio files may form a test set to test the trained ASR module 206.

[0054] The ASR module 206 may be trained and/or re-trained based on the review module 208. In some embodiments, the ASR module 206 is configured to automatically identify patterns in an input audio file comprising one or more speech waveforms. Patterns that could be detected from speech may include the speaker's identity (used for voice verification), language, emotion and the textual transcription of the spoken utterance.

[0055] In some embodiments, the ASR module 206 may include a Machine Learning (ML) algorithm or Artificial Intelligence (AI) algorithm to process human speech into readable text. In some embodiments, the ML/AI algorithm may be trained using unsupervised learning, supervised learning, and/or hybrid training methodology. In some embodiments, the ASR module 206 may include one or more neural network algorithms. In some embodiments, the ASR module 206 may be modelled as a pattern classifier and configured or programmed to process any input audio files using one or more of the following known techniques: feature extraction, segmentation, embedding, clustering, mapping/classification, and refinement. An example of the AI/ML algorithm will be a deep learning convolutional neural network.

[0056] The review module 208 is configured to review the output text data (transcribed data) of the ASR module 206. In some embodiments, the review module 208 includes a check on the output text data and identifying words having relatively lower precision/recall rates. In some embodiments, the review module 208 may include obtaining/receiving manual intervention and feedback, such that if users feel that some words have not been identified well, feedback may be sent to the selector module 202 to include more of the words that have not been identified well. The selector module 202 may then be configured to customize the generation of the text dataset such that the words that have not been identified well may be selected for re-training.

[0057] In some embodiments, one or more of the modules 202, 204, 206 and 208 may be combined or integrated as one single module. For example, the functions associated with modules 202, 204, 206 and 208 may be implemented by one single computer processor, or a group of computer processors.

[0058] FIG. 3 shows another embodiment of a system 300 for building a dataset for automatic speech recognition. The system 300 implements a cross-validation model to improve the quality and robustness of pseudo labels that can be assigned to the audio files. Two sets of collected data 302a, 302b, in the form of audio files, may be input into a trained automatic speech recognition module 306a, 306b respectively. The output data 308a of the speech

recognition module 306a may then be fed with the input data 302b into an inference model 310a to perform inference on the output data 308a based on input data 302b. The output data 308b of the speech recognition module 306b may then be fed with the input data 302a into an inference model 310b to perform inference on the output data 308b based on input data 302a.

[0059] The output from the inference model 310a and 310b may be compared, and any wrong labels may be corrected and/or removed. For example, if the outputs from 310a and 310b are different or wrongly identified based on a comparison with a selected set of pseudo-labels identified to be always true or near true, the wrong labels may be corrected and/or removed.

[0060] FIG. 4 and FIG. 5 show two applications of the trained ASR module of the present disclosure.

[0061] FIG. 4 shows an application of the trained ASR module in the form of a post-call analysis system 400 between a customer service officer (also referred to as an agent), and a customer. A trained ASR module 450 may be deployed or implemented as a complementary service to a known conversation recording system. The ASR module 450 may be configured to transcribe the conversation into text on a real-time or near real-time basis. In FIG. 4, two scenarios are shown, one without the ASR module 450 (marked A), and the other with the ASR module 450 (marked B).

[0062] In scenario A, a supervisory personnel, for example a manager will be required to listen to the conversation recording system to identity the context and issues associated with the call. The manager will be required to call the customer for resolution if necessary and if the customer mentioned something else that happened during the call with the agent, the manager will be required to listen to the conversation recording system to re-confirm if anything has been missed out.

[0063] In scenario B, the ASR module 450 performs a transcript of the conversation based on the audio file obtained from the known conversation recording system. The manager may read the transcript and can refer to it while calling the customer for resolution, and should the customer mentioned something else that happened during the call with the agent, the manager may simply take a quick glance at the transcript. Compared to scenario A, scenario B enables a manager to concurrently call back a customer and identify key issues by looking at the transcript.

[0064] FIG. 5 shows another application 500 of a transcription of conversation between a service provider (e.g. vehicle on-demand driver) and a user (e.g. passenger) vehicle in-trip audio recording. A trained ASR module 550 may be deployed or implemented as a complementary service to an in-vehicle conversation recording system. The ASR module 550 may be configured to transcribe the conversation into text on a real-time or near real-time basis. In FIG. 5, two scenarios are shown, one without the ASR module 550 (marked A), and the other with the ASR module 550 (marked B). The service provider may submit an in-trip recording if the trip ended with conflict(s) for further investigation by an authorized party.

[0065] In scenario A, a personnel, for example an administrator, may be required to listen to the conversation recording system to identify the conflict associated with the recording. The administrator will be required to transcribe the conversation and produce a report for onward submission to the next personnel in charge, for example a manager. There may be at least one round of clarification between the administrator and the manager to finalize the report.

[0066] In scenario B, the ASR module 550 performs a transcript of the conversation based on the audio file obtained from the in-vehicle recording system. The administrator may read the transcript while preparing the report and refer to the audio recording when required. Likewise, the transcript is sent to the manager together with the audio recording for completeness. Compared to scenario A, scenario B provides for more efficient time management associated with generating a conflict report.

[0067] The method of FIG. 1 is for example carried out by a server computer as illustrated in FIG. 6.

[0068] FIG. 6 shows a server computer 600 according to an embodiment.

[0069] The server computer 600 includes a communication interface 601 (e.g. configured to receive interaction data, i.e. information about interactions). The server computer 600 further includes a processing unit 602 and a memory 603. The memory 603 may be used by the processing unit 602 to store, for example, data to be processed, such as generated text files and audio files. The server computer is configured to perform the method of FIG. 1.

[0070] The methods described herein may be performed and the various processing or computation units and the devices and computing entities described herein may be implemented by one or more circuits. In an embodiment, a "circuit" may be understood as any kind of a logic implementing entity, which may be hardware, software, firmware, or any combination thereof. Thus, in an embodiment, a "circuit" may be a hard-wired logic circuit or

a programmable logic circuit such as a programmable processor, e.g. a microprocessor. A "circuit" may also be software being implemented or executed by a processor, e.g. any kind of computer program, e.g. a computer program using a virtual machine code. Any other kind of implementation of the respective functions which are described herein may also be understood as a "circuit" in accordance with an alternative embodiment.

[0071] While the disclosure has been particularly shown and described with reference to specific embodiments, it should be understood by those skilled in the art that various changes in form and detail may be made therein without departing from the spirit and scope of the disclosure as defined by the appended claims. The scope of the disclosure is thus indicated by the appended claims and all changes which come within the meaning and range of equivalency of the claims are therefore intended to be embraced.

CLAIMS

1. A computer-implemented method for building at least one dataset for automatic speech recognition, the method comprising
 - generating an input dataset;
 - obtaining an audio dataset associated with the input dataset;
 - verifying the audio dataset using an audio verification module; assigning at least one pseudo-label to the verified audio dataset;
 - storing the at least one pseudo-labelled audio dataset in a training database;
 - repeating the above steps until a pre-determined number of pseudo-labelled audio dataset have been stored in the training database; and
 - training an automatic speech recognition module based on the pre-determined number of pseudo-labelled audio dataset.
2. The computer-implemented method of claim 1, wherein the input dataset comprises a text dataset, and generating the text dataset comprises introducing at least one of a relatively uncommon vocabulary and a relatively uncommon language in the text dataset.
3. The computer-implemented method of claim 2, further comprises comparing the text dataset for speaker verification with an output transcription file of the automatic speech recognition module, and identifying at least one unsatisfactory result from the comparison.
4. The computer-implemented method of claim 3, wherein identifying the at least one unsatisfactory trained result includes determining a precision rate and/or a recall rate associated with one or more words in the text dataset.
5. The computer-implemented method of claim 3 or 4, further comprising selecting a speaker from a user group historically associated with accurate text-dependent speaker verification, and assigning the at least one pseudo-label to the audio dataset generated by the speaker without verification.

6. The computer-implemented method of any one of the preceding claims, further comprising generating an output transcription file by the automatic speech recognition module.
7. The computer-implemented method of claim 1, wherein the input dataset comprises a first text dataset and a second text dataset, resulting in a respective first output transcription file and a second output transcription file, wherein the first text dataset and the second output transcription file is configured to be input to a first inference module, and the second text dataset and the first output transcription file is configured to be input to a second inference module.
8. The computer-implemented method of claim 7, wherein the output of first inference module and the output of the second inference module are compared.
9. A device for building a dataset for automatic speech recognition, the device comprises a processor configured to repeatedly:
 - generate an input dataset for audio verification;
 - obtain an audio dataset based on the generated input dataset;
 - verify the audio dataset using an audio verification module; assign at least one pseudo-label to the verified audio dataset;
 - store the at least one pseudo-labelled audio dataset in a training database;until a pre-determined number of pseudo-labelled audio datasets have been stored in the training database.
10. The device of claim 9, wherein the input dataset is a text dataset, and the processor is configured to introduce at least one of a relatively uncommon vocabulary and a relatively uncommon language in the generation of the text dataset.
11. The device of claim 10, wherein the processor is configured to compare the text dataset for speaker verification with an output transcription file of the automatic speech recognition module, and identify at least one unsatisfactory result from the comparison.

12. The device of claim 11, wherein the processor is configured to identify the at least one unsatisfactory trained result based on a precision rate and/or a recall rate associated with one or more words in the output transcription file.

13. The device of claim 11 or 12, wherein the processor is configured to select a speaker from a user group historically associated with accurate speaker verification, and the processor is further configured to assign the at least one pseudo-label to the audio dataset generated by the speaker, without verification.

14. The device of claim 11, wherein the processor is configured to generate the text dataset based on feedback of the at least one unsatisfactory result.

15. The device of claim 9, wherein the input dataset comprises a first text dataset and a second text dataset, resulting in a first output transcription file and a second output transcription file respectively, wherein the first text dataset and the second output transcription file is configured to be input to a first inference module, and the second text dataset and the first output transcription file is configured to be input to a second inference module.

16. The device of claim 15, wherein the output of first inference module and the output of the second inference module are compared.

17. A non-transitory computer-readable storage medium comprising instructions, which, when executed by one or more processors, cause the execution of the method for building a dataset for automatic speech recognition according to any one of claims 1 to 8.

18. A data processing device configured to perform the method of any one of claims 1 to 8.

19. A computer executable code comprising instructions building a dataset for automatic speech recognition according to any one of claims 1 to 8.

20. An automatic speech recognition module trained by the method of any one of claims 1 to 8.